

Statistical Learning under Nonstationary Mixing Processes

Steve Hanneke*

Tommi Jaakkola[†]

Liu Yang[‡]

Abstract

We study a special case of the problem of statistical learning without the i.i.d. assumption. Specifically, we suppose a learning method is presented with a sequence of data points, and required to make a prediction (e.g., a classification) for each one, and can then observe the loss incurred by this prediction. We go beyond traditional analyses, which have focused on stationary mixing processes or nonstationary product processes, by combining these two relaxations to allow nonstationary mixing processes. We are particularly interested in the case of β -mixing processes, with the sum of changes in marginal distributions growing sublinearly in the number of samples. Under these conditions, we propose a learning method, and establish that for bounded VC subgraph classes, the cumulative excess risk grows sublinearly in the number of predictions, at a quantified rate.

Keywords: Statistical Learning Theory; Nonstationary Processes; VC Theory.

AMS MSC 2010: 68T05; 68Q32; 62M99; 60G99.

1 Introduction

Our setting is that of stream-based prediction. At each time t , we are given access to data points from times 1 through $t - 1$, and are required to produce a predictor f_t , which is then evaluated on a new data point at time t . We study this in the *general learning setting* of [28, 29], which represents the learning objective as an abstract optimization problem. As an example, in the special case of classification, given access to pairs $(x_1, y_1), \dots, (x_{t-1}, y_{t-1})$, we would be tasked with producing a function mapping an observed point x_t to a classification \hat{y}_t , and we would be evaluated on whether $\hat{y}_t \neq y_t$ (called a *mistake*). We are then interested in characterizing the rate of growth of the cumulative number of mistakes, as we repeat this for increasing values of t .

To study this problem, we suppose the sequence of observations are stochastic, subject to some restrictions on their distribution. Several such restrictions are possible. For instance, the most-common assumption used in the vast majority of the statistical learning literature is that the data are independent and identically distributed (i.i.d.). However, some efforts to relax this assumption have also been explored. There are essentially two main threads of work toward relaxing this assumption: relaxing the independence assumption while maintaining the assumption of identical distributions (or stationarity), or relaxing the assumption of identical distributions while maintaining the independence assumption. In the present work, we are interested in relaxing these assumptions jointly. Before getting into the details, let us first briefly review these two threads of the literature.

Most of the literature on relaxations of the independence assumption focuses on *stationary mixing* processes. At the extreme of this branch, the work of [1] reveals that any VC class admits a uniform law of large numbers under stationary ergodic processes. In particular, this implies that the method of *empirical risk minimization* approaches

*E-mail: steve.hanneke@gmail.com.

[†]Massachusetts Institute of Technology. E-mail: tommy@csail.mit.edu.

[‡]IBM T. J. Watson Research Center. E-mail: yangli@us.ibm.com.

excess risk zero in the limit. However, one cannot establish *rates* of convergence under such general conditions as ergodicity. To establish such rates, other works have therefore introduced stronger conditions, such as the β -mixing condition. Specifically, [34] has proven asymptotic rates of uniform convergence for VC classes under stationary β -mixing processes. One implication of this result is an asymptotic rate of convergence for the excess risk of empirical risk minimization. Other works have established rates of convergence for the excess risk of empirical risk minimization and other learning methods, under related mixing conditions, including α -mixing [31], η -mixing [17], and ϕ -mixing [31], all under the stationarity assumption.

The other primary direction in the study of the risk of learning methods under relaxations of the i.i.d. assumption preserves the independence assumption, while allowing the marginal distributions to *drift* over time. This thread in the literature has focused on the specific setting of binary classification. Specifically, [19, 14, 15, 5, 6, 9] study a setting in which the marginal distribution of the data point at time t has total variation distance from that of the data point at time $t + 1$ at most a given upper bound, called the *drift rate* (see also related work by [3, 11, 4, 33, 22]). The data points are still assumed to be independent. The recent works of [12, 22] further explore this problem (in a formulation more-closely paralleling that studied here). In this setting, the learning method produces a sequence of predictors (e.g., classifiers), where the method for choosing the predictor at time t may depend on all of the data up to time $t - 1$. The results in these works are expressible as bounds on the risk at each time t (or sometimes averaged over time), as a function of t and the rates of drift of the marginal distributions.

The paper of [22] also studies a refinement of the notion of “drift” compared to the earlier works, such as [5, 6]. Specifically, rather than measuring the difference between the next and previous distributions by the total variation distance, they instead use a notion of “discrepancy” that depends directly on the function class being used for learning. This discrepancy is sometimes significantly smaller than the total variation distance, yet plays an analogous role in the bounds of [22] as the total variation distance plays in the bounds of [15, 6]. To allow for this refined notion of drift, our arguments below are phrased generally enough that they can be applied with either notion of drift (discrepancy or total variation).

In recent work, [18] discusses the problem of learning from non-stationary mixing processes. They derive interesting results bounding the risk at some future time in terms of the empirical risk on all observed data, with clear implications for the performance of methods such as empirical risk minimization. The nature of the results in that work are somewhat different from our results below. However, the spirit of the analysis is similar in many places, and one can conceivably convert some of those results into a more-closely related form with a bit of additional effort.

One significant point of divergence between the present work and that of [18], and indeed all of the above works on product processes (aside from certain special cases discussed by [12]), is that in the general case, these works require access to the sequence of magnitudes of drift of the distribution, or a constant upper bound thereon. The sequence of drift magnitudes is a substantial number of variables to assume we have access to (linear in the number of data points), and relying only on a constant upper bound precludes the possibility of sublinear growth of the cumulative excess risk [15, 12]. The notion of discrepancy studied by [22, 18] (see below) can sometimes be estimated from data, but only under significant further restrictions on the process. In contrast, in the present work, we merely assume an asymptotic bound on the rate of growth of the cumulative amount of drift. Our learning method then depends only on the single parameter that this asymptotic growth rate is described in terms of, and we show that this is enough to achieve sublinear growth of the cumulative excess risk, without

needing access to the sequence of drift rates or additional restrictions on the process. For completeness, we also briefly discuss the case where the drift rates are known, in Section 3.

The present work studies learning under general nonstationary processes, under a condition that allows us to extend the ideas from the above-described literature on learning from product processes with slowly-drifting marginal distributions. Specifically, we replace the independence condition with a β -mixing condition. In addition to this, we suppose that the sum of distances between marginal distributions at adjacent time steps grows only sublinearly (note that this does *not* require that the sequence of distributions be converging). Our objective is then to propose a prediction strategy (for producing the f_t function), and to characterize the rate of growth of the cumulative excess risk over time. The excess risks are calculated relative to the sequence of *a priori* optimal predictors among functions in a given function class. In particular, for any bounded VC subgraph class, we establish a rate of growth of the cumulative excess risk that is *sublinear* in the number of predictions made.

1.1 Definitions and Summary of Main Result

To formalize this setting, we adopt the abstract perspective of the *general learning setting* of [28, 29]. Specifically, fix a measurable space $(\mathcal{Z}, \mathcal{Z})$ and a *function class* \mathcal{F} of measurable functions $f : \mathcal{Z} \rightarrow [0, 1]$. For instance, in the special case of classification, \mathcal{Z} would be a set of (x, y) pairs, and \mathcal{F} would be a set of functions $f_h((x, y)) = \mathbb{1}[h(x) \neq y]$, where h ranges over a set \mathcal{H} of functions (known as the hypothesis class); see [16, 25] for many other examples. In the general learning setting, the aim of a learning algorithm is to identify a function $f \in \mathcal{F}$ with a relatively small average value, where the average is taken with respect to some unknown probability measure on \mathcal{Z} (as discussed in more detail below). For instance, in the classification setting described above, this average value corresponds to the probability that h makes a “mistake” in predicting the value of y from x .

For simplicity, to avoid the common measurability issues arising in empirical process theory, we will suppose \mathcal{F} is such that the events involved in the proofs below are all measurable (for instance, this is certainly the case if \mathcal{F} is countable; see [27] for other sufficient conditions). Let d denote the pseudo-dimension of \mathcal{F} [23, 24, 13, 2]: that is, d is the largest $k \in \mathbb{N} \cup \{0\}$ such that $\exists (z_1, w_1), \dots, (z_k, w_k) \in \mathcal{Z} \times \mathbb{R}$ with $|\{(\mathbb{1}[f(z_1) \leq w_1], \dots, \mathbb{1}[f(z_k) \leq w_k]) : f \in \mathcal{F}\}| = 2^k$, or is ∞ if no such largest k exists. Throughout this article, we suppose $1 \leq d < \infty$ (so that \mathcal{F} is a VC Subgraph class).

We suppose there is a sequence of \mathcal{Z} -valued random variables Z_1, Z_2, \dots , called the *data points*, and for each $t \in \mathbb{N}$, we denote by P_t the marginal distribution of the random variable Z_t . Also, generally, for any random variable X , we denote by \mathbb{P}_X the distribution of X (i.e., $\mathbb{P}_X(\cdot) = \mathbb{P}(X^{-1}(\cdot))$). For any probability measures P, Q on a measurable space (Ω, \mathcal{B}) , we denote by $\|P - Q\| = \sup_{A \in \mathcal{B}} P(A) - Q(A)$ the total variation distance between P and Q . Additionally, for probability measures P, Q on the measurable space $(\mathcal{Z}, \mathcal{Z})$, we denote by

$$\rho(P, Q) = \sup_{f \in \mathcal{F}} |\mathbb{E}_{Z \sim P}[f(Z)] - \mathbb{E}_{Z \sim Q}[f(Z)]|,$$

a general notion of *discrepancy* introduced by [20, 22]. We use ρ below to quantify the magnitude of change in the marginal distribution of Z_{t+1} compared to Z_t . Note that, since every $f \in \mathcal{F}$ is uniformly bounded in $[0, 1]$, we clearly have

$$\rho(P, Q) \leq \|P - Q\|.$$

Indeed, readers more comfortable with the familiar total variation distance may feel free to replace $\rho(P, Q)$ with $\|P - Q\|$ in all contexts below, and the results and proofs will

remain valid without any further modifications. However, one can construct scenarios in which $\rho(P, Q)$ provides a much smaller value, and generally $\rho(P, Q)$ appears to be more relevant to the learning setting than is the total variation distance. For each $t \geq 2$, let $\Delta_t \in [0, 1]$ be a value satisfying

$$\rho(P_t, P_{t-1}) \leq \Delta_t. \quad (1.1)$$

For completeness, also define $\Delta_1 = 0$.

To obtain nontrivial results, we are interested in restricting the family of processes. Specifically, for our main result below (Theorem 1.1), we suppose

$$\sum_{t=1}^T \Delta_t = O(T^\alpha), \quad (1.2)$$

for a given value $\alpha \in [0, 1)$. Note that this does not require that the sequence of distributions be converging, only that its average rate of change slows over time. We additionally adopt the standard definition of β -mixing, defined as follows. Following [8] and [34], for each $k \in \mathbb{N}$, define

$$\beta_k = \frac{1}{2} \sup \left\{ \sum_{i=1}^I \sum_{j=1}^J |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)| : \{A_i\}_i \in \Pi_\ell, \{B_j\}_j \in \Pi'_{\ell+k}, \ell \geq 1 \right\},$$

where Π_ℓ is defined as the set of $\sigma(\{Z_1, \dots, Z_\ell\})$ -measurable finite partitions, and $\Pi'_{\ell+k}$ is defined as the set of $\sigma(\{Z_{\ell+k}, Z_{\ell+k+1}, \dots\})$ -measurable finite partitions. Then we suppose

$$\beta_k = O(k^{-r}), \quad (1.3)$$

for some $r \in (0, \infty)$.

Under the assumptions (1.2) and (1.3), we propose a learning method, specified as follows. Let \hat{f}_1 be arbitrary. For each $t \in \mathbb{N} \setminus \{1\}$, let

$$k_t = \left\lceil t^{(1-\alpha)\frac{3}{3+4r}} \right\rceil \wedge (t-1)$$

and

$$m_t = \left\lceil t^{(1-\alpha)\frac{3+2r}{3+4r}} \right\rceil \wedge (t-1),$$

and choose as a predictor at time t a function¹

$$\hat{f}_t = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{s=1}^{\lfloor m_t/k_t \rfloor} f(Z_{t-sk_t}). \quad (1.4)$$

For completeness, \hat{f}_1 can be defined as an arbitrary function in \mathcal{F} . For \hat{f}_t chosen in this way, we prove the following theorem.

Theorem 1.1. *If (1.2) and (1.3) are satisfied, then*

$$\sum_{t=1}^T \mathbb{E} [\hat{f}_t(Z_t)] - \sum_{t=1}^T \inf_{f \in \mathcal{F}} \mathbb{E} [f(Z_t)] = O \left(T^{\alpha + (1-\alpha)\frac{3+3r}{3+4r}} \right).$$

In particular, note that the expression on the right hand side grows *sublinearly* in T . To prove this theorem, we first provide two key lemmas from the literature, after which we present the proof of Theorem 1.1 below. Following this, in Section 3, we conclude the paper by establishing *finite-sample* bounds, and other specialized results, in the special case of *product* processes; this effectively extends to the general learning setting results established by [5, 6] for binary classification, while also expressing the results in a more general form that allows for a time-varying drift rate.

¹For simplicity, we suppose the minimum is actually *achieved* by some $f \in \mathcal{F}$. To handle the general case, all of the results continue to hold, with only minor technical changes to the proofs, if we instead choose $\hat{f}_t \in \mathcal{F}$ with $\sum_{s=1}^{\lfloor m_t/k_t \rfloor} \hat{f}_t(Z_{t-sk_t})$ sufficiently close to $\inf_{f \in \mathcal{F}} \sum_{s=1}^{\lfloor m_t/k_t \rfloor} f(Z_{t-sk_t})$.

2 Proof of Theorem 1.1

The following lemma is a well-known result on β -mixing processes, from [32, 10] (see also Theorem 2.1 of [31] or Corollary 2.7 of [34]).

Lemma 2.1. *For any $t, n, k \in \mathbb{N}$,*

$$\left\| \mathbb{P}_{\{Z_{(j-1)k+t}\}_{j=1}^n} - \left(\times_{j=1}^n P_{(j-1)k+t} \right) \right\| \leq (n-1)\beta_k.$$

Additionally, we use the following well-known result of [26] (which refines earlier results of [30, 13]).

Lemma 2.2. *There exists a universal constant $c \in [1, \infty)$ such that, for any independent \mathcal{Z} -valued random variables Z'_1, \dots, Z'_m ,*

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{t=1}^m (f(Z'_t) - \mathbb{E}[f(Z'_t)]) \right| \right] \leq c \sqrt{\frac{d}{m}}.$$

Specifically, this result is obtained by integrating the tail bound in Theorem 1.3 of [26] (see also [27], Theorem 2.14.9); see e.g., [27], for verification that VC subgraph classes indeed satisfy the conditions on \mathcal{F} in the more-general original statement of [26]. While the original proof of this result by [26] discussed only i.i.d. random variables, the proof in fact implies this result, which only assumes independence. For completeness, we include a brief proof in Appendix A.

With these lemmas in hand, we are ready to present the proof of Theorem 1.1.

Proof of Theorem 1.1. Let Z'_1, Z'_2, \dots denote a sequence of independent random variables, also independent from $\{Z_i\}_{i \in \mathbb{N}}$, and with each $Z'_i \sim P_i$. Fix any $t \in \mathbb{N} \setminus \{1\}$. Since \hat{f}_t depends only on Z_1, \dots, Z_{t-k_t} , it follows immediately from the definition of β_{k_t} (see [34], Lemma 2.6) that

$$\left\| \mathbb{P}_{(\hat{f}_t, Z_t)} - \mathbb{P}_{(\hat{f}_t, Z'_t)} \right\| = \left\| \mathbb{P}_{(\hat{f}_t, Z_t)} - \mathbb{P}_{\hat{f}_t} \times \mathbb{P}_{Z_t} \right\| \leq \beta_{k_t}.$$

In particular, this implies

$$\mathbb{E} [\hat{f}_t(Z_t)] \leq \mathbb{E} [\hat{f}_t(Z'_t)] + \beta_{k_t}.$$

Additionally, since $\rho(P_{t-ik_t}, P_t) \leq 1 \wedge \sum_{q=t-ik_t}^{t-1} \Delta_{q+1} \leq 1 \wedge \sum_{q=t-m_t}^{t-1} \Delta_{q+1}$ for $1 \leq i \leq \lfloor m_t/k_t \rfloor$, and every Z'_j is independent of \hat{f}_t , we have that

$$\begin{aligned} \mathbb{E} [\hat{f}_t(Z'_t)] &= \mathbb{E} [\mathbb{E} [\hat{f}_t(Z'_t) | \hat{f}_t]] \\ &\leq \mathbb{E} \left[\frac{1}{\lfloor m_t/k_t \rfloor} \sum_{i=1}^{\lfloor m_t/k_t \rfloor} \mathbb{E} [\hat{f}_t(Z'_{t-ik_t}) | \hat{f}_t] \right] + 1 \wedge \sum_{q=t-m_t}^{t-1} \Delta_{q+1}. \end{aligned}$$

Furthermore,

$$\begin{aligned} &\mathbb{E} \left[\frac{1}{\lfloor m_t/k_t \rfloor} \sum_{i=1}^{\lfloor m_t/k_t \rfloor} \mathbb{E} [\hat{f}_t(Z'_{t-ik_t}) | \hat{f}_t] \right] \\ &\leq \mathbb{E} \left[\frac{1}{\lfloor m_t/k_t \rfloor} \sum_{i=1}^{\lfloor m_t/k_t \rfloor} \hat{f}_t(Z_{t-ik_t}) \right] + \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{\lfloor m_t/k_t \rfloor} \sum_{i=1}^{\lfloor m_t/k_t \rfloor} (\mathbb{E}[f(Z'_{t-ik_t})] - f(Z_{t-ik_t})) \right| \right]. \end{aligned} \tag{2.1}$$

Now let us bound each term in (2.1) separately. First, we have that

$$\begin{aligned} \mathbb{E} \left[\frac{1}{\lfloor m_t/k_t \rfloor} \sum_{i=1}^{\lfloor m_t/k_t \rfloor} \hat{f}_t(Z_{t-ik_t}) \right] &= \mathbb{E} \left[\inf_{f \in \mathcal{F}} \frac{1}{\lfloor m_t/k_t \rfloor} \sum_{i=1}^{\lfloor m_t/k_t \rfloor} f(Z_{t-ik_t}) \right] \\ &\leq \inf_{f \in \mathcal{F}} \frac{1}{\lfloor m_t/k_t \rfloor} \sum_{i=1}^{\lfloor m_t/k_t \rfloor} \mathbb{E}[f(Z_{t-ik_t})] \leq \inf_{f \in \mathcal{F}} \mathbb{E}[f(Z_t)] + 1 \wedge \sum_{q=t-m_t}^{t-1} \Delta_{q+1}. \end{aligned}$$

Next, Lemma 2.1 implies

$$\begin{aligned} &\mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{\lfloor m_t/k_t \rfloor} \sum_{i=1}^{\lfloor m_t/k_t \rfloor} (\mathbb{E}[f(Z'_{t-ik_t})] - f(Z_{t-ik_t})) \right| \right] \\ &\leq \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{\lfloor m_t/k_t \rfloor} \sum_{i=1}^{\lfloor m_t/k_t \rfloor} (\mathbb{E}[f(Z'_{t-ik_t})] - f(Z'_{t-ik_t})) \right| \right] + (\lfloor m_t/k_t \rfloor - 1) \beta_{k_t}. \end{aligned}$$

Furthermore, Lemma 2.2 implies

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{\lfloor m_t/k_t \rfloor} \sum_{i=1}^{\lfloor m_t/k_t \rfloor} (\mathbb{E}[f(Z'_{t-ik_t})] - f(Z'_{t-ik_t})) \right| \right] \leq c \sqrt{\frac{d}{\lfloor m_t/k_t \rfloor}}.$$

Together, we have that (2.1) is at most

$$\inf_{f \in \mathcal{F}} \mathbb{E}[f(Z_t)] + 1 \wedge \sum_{q=t-m_t}^{t-1} \Delta_{q+1} + c \sqrt{\frac{d}{\lfloor m_t/k_t \rfloor}} + (\lfloor m_t/k_t \rfloor - 1) \beta_{k_t}.$$

Altogether, we have established that

$$\mathbb{E} [\hat{f}_t(Z_t)] \leq \inf_{f \in \mathcal{F}} \mathbb{E}[f(Z_t)] + 2 \left(1 \wedge \sum_{q=t-m_t}^{t-1} \Delta_{q+1} \right) + c \sqrt{\frac{d}{\lfloor m_t/k_t \rfloor}} + \lfloor m_t/k_t \rfloor \beta_{k_t}. \quad (2.2)$$

Therefore,

$$\begin{aligned} &\sum_{t=1}^T \mathbb{E} [\hat{f}_t(Z_t)] - \sum_{t=1}^T \inf_{f \in \mathcal{F}} \mathbb{E}[f(Z_t)] \\ &\leq 1 + \left(\sum_{t=2}^T 2 \left(1 \wedge \sum_{q=t-m_t}^{t-1} \Delta_{q+1} \right) \right) + \left(\sum_{t=2}^T c \sqrt{\frac{d}{\lfloor m_t/k_t \rfloor}} \right) + \left(\sum_{t=2}^T \lfloor m_t/k_t \rfloor \beta_{k_t} \right). \quad (2.3) \end{aligned}$$

All that remains is to bound each of these three terms on the right hand side of (2.3). First, note that

$$\sum_{t=1}^T t^{-(1-\alpha)\frac{r}{3+4r}} = O \left(1 + \int_1^T t^{-(1-\alpha)\frac{r}{3+4r}} dt \right) = O \left(T^{\alpha+(1-\alpha)\frac{3+3r}{3+4r}} \right).$$

Thus, we have that

$$\sum_{t=2}^T c \sqrt{\frac{d}{\lfloor m_t/k_t \rfloor}} = O \left(\sum_{t=1}^T t^{-(1-\alpha)\frac{r}{3+4r}} \right) = O \left(T^{\alpha+(1-\alpha)\frac{3+3r}{3+4r}} \right). \quad (2.4)$$

Also, we have

$$\sum_{t=2}^T \lfloor m_t/k_t \rfloor \beta_{k_t} = O \left(\sum_{t=2}^T m_t/k_t^{1+r} \right) = O \left(\sum_{t=1}^T t^{-(1-\alpha)\frac{r}{3+4r}} \right) = O \left(T^{\alpha+(1-\alpha)\frac{3+3r}{3+4r}} \right). \quad (2.5)$$

The final term, $\sum_{t=2}^T 2 \left(1 \wedge \sum_{q=t-m_t}^{t-1} \Delta_{q+1} \right)$, requires slightly more work to bound. Define $\delta_t = t^{-(1-\alpha)\frac{r}{3+4r}}$, and let $I_t = \mathbb{1} \left[\sum_{q=t-m_t}^{t-1} \Delta_{q+1} > \delta_t \right]$. Then we have $1 \wedge \sum_{q=t-m_t}^{t-1} \Delta_{q+1} \leq \delta_t + I_t$, so that

$$\sum_{t=2}^T \left(1 \wedge \sum_{q=t-m_t}^{t-1} \Delta_{q+1} \right) \leq \left(\sum_{t=2}^T \delta_t \right) + \left(\sum_{t=2}^T I_t \right) = O \left(T^{\alpha+(1-\alpha)\frac{3+3r}{3+4r}} \right) + \sum_{t=2}^T I_t. \quad (2.6)$$

Now we claim that

$$\sum_{t=2}^T \frac{1}{m_t} \sum_{q=t-m_t}^{t-1} \Delta_{q+1} = O \left(\sum_{t=1}^T \Delta_t \right).$$

To see this, note that the smallest value of t for which $q \in \{t - m_t + 1, \dots, t\}$ has $q = t$, while the largest value of t for which $q \in \{t - m_t + 1, \dots, t\}$ has $q = t - m_t + 1$. In particular, this means that for this largest value of t , $2q = 2t \left(1 - \frac{m_t-1}{t} \right) \geq 2t \left(1 - \frac{m_q}{q} \right)$, which for any sufficiently large q , is greater than t : that is, $2q > t$. Therefore, together with monotonicity of m_s , for any sufficiently large q , this largest value of t with $q \in \{t - m_t + 1, \dots, t\}$ satisfies $t = q + m_t - 1 \leq q + m_{2q} - 1$. Furthermore, sublinearity of m_t implies that every value of q has only a finite number of values t for which $q \in \{t - m_t + 1, \dots, t\}$. Therefore, there exists a T -independent value $v_0 \in (0, \infty)$ (to account for all of the Δ_q terms with insufficiently large q values for the above argument) such that

$$\begin{aligned} \sum_{t=2}^T \frac{1}{m_t} \sum_{q=t-m_t+1}^t \Delta_q &\leq v_0 + \sum_{q=2}^T \sum_{t=q}^{q+m_{2q}-1} \frac{\Delta_q}{m_t} \leq v_0 + \sum_{q=2}^T \frac{m_{2q}}{m_q} \Delta_q \\ &\leq v_0 + \sum_{q=2}^T 4\Delta_q = O \left(\sum_{q=1}^T \Delta_q \right). \end{aligned}$$

Furthermore, by (1.2), this is $O(T^\alpha)$. Additionally note that

$$\sum_{t=2}^T \frac{1}{m_t} \sum_{q=t-m_t}^{t-1} \Delta_{q+1} \geq \sum_{t=2}^T \frac{\delta_t}{m_t} I_t \geq \frac{\delta_T}{m_T} \sum_{t=2}^T I_t.$$

Together, these two facts imply that

$$\sum_{t=2}^T I_t \leq O \left(\frac{m_T}{\delta_T} \sum_{t=1}^T \Delta_t \right) = O \left(T^{\alpha+(1-\alpha)\frac{3+3r}{3+4r}} \right).$$

Plugging this into (2.6), we have that

$$\sum_{t=2}^T \left(1 \wedge \sum_{q=t-m_t}^{t-1} \Delta_{q+1} \right) = O \left(T^{\alpha+(1-\alpha)\frac{3+3r}{3+4r}} \right).$$

Along with (2.3), (2.4), and (2.5), we have established that

$$\sum_{t=1}^T \mathbb{E} \left[\hat{f}_t(Z_t) \right] - \sum_{t=1}^T \inf_{f \in \mathcal{F}} \mathbb{E} [f(Z_t)] = O \left(T^{\alpha+(1-\alpha)\frac{3+3r}{3+4r}} \right),$$

which completes the proof. \square

3 Product Processes

In this section, unlike above, we suppose the algorithm has direct access to the Δ_t sequence. Our objective is then to derive more-explicit (non-asymptotic) bounds under the assumption that $\{Z_t\}_{t=1}^\infty$ is a product process. The results here are already known in the special case of binary classification, in the case that Δ_t is bounded by a t -invariant *constant* for all t [6]. Thus, this section represents a generalization of these classic results to the general learning setting, and to general time-varying drift rates. That said, we note that the results here would also readily follow from the classic analysis of [6] and the more-recent work of [22], with only minor additional work to apply those results to a recent history of data points trailing the prediction time t .

Throughout this section, for any functions $f, g : A \rightarrow [0, \infty)$, for any set A , we write $f(a) \lesssim g(a)$ to express the claim that there exists a numerical constant $c \in (0, \infty)$ such that $f(a) \leq cg(a)$ for all $a \in A$; this allows us to express non-asymptotic bounds (in terms of T , d , and the Δ_t sequence), without concerning ourselves with precise numerical constant factors. For each $t \in \mathbb{N} \setminus \{1\}$, define

$$\tilde{m}_t = \operatorname{argmin}_{m \in \{1, \dots, t-1\}} \left(\sum_{q=t-m}^{t-1} \Delta_{q+1} + \sqrt{\frac{d}{m}} \right)$$

and

$$\tilde{f}_t = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{s=t-\tilde{m}_t}^{t-1} f(Z_s).$$

For completeness, define \tilde{f}_1 as an arbitrary element of \mathcal{F} .

Theorem 3.1. *If $\{Z_t\}_{t=1}^\infty$ is a product process, then for $T \in \mathbb{N} \setminus \{1\}$,*

$$\sum_{t=1}^T \mathbb{E} [\tilde{f}_t(Z_t)] - \sum_{t=1}^T \inf_{f \in \mathcal{F}} \mathbb{E} [f(Z_t)] \lesssim \sum_{t=2}^T \min_{m \in \{1, \dots, t-1\}} \left(\sum_{q=t-m}^{t-1} \Delta_{q+1} + \sqrt{\frac{d}{m}} \right).$$

Proof. We begin by noting that, in the proof of Theorem 1.1, the argument leading to (2.3) in fact more generally holds for any process $\{Z_t\}_{t \in \mathbb{N}}$ (regardless of whether (1.2) and (1.3) are satisfied), and for any sequence \hat{f}_t defined as in (1.4), where the values $m_t, k_t \in \mathbb{N}$ can be specified *arbitrarily*, subject to $k_t \leq m_t \leq t-1$. In particular, substituting $k_t = 1$ and $m_t = \tilde{m}_t$, the corresponding \hat{f}_t from (1.4) is precisely \tilde{f}_t . Then since $\beta_1 = 0$ for product processes, (2.3) implies

$$\begin{aligned} \sum_{t=1}^T \mathbb{E} [\tilde{f}_t(Z_t)] - \sum_{t=1}^T \inf_{f \in \mathcal{F}} \mathbb{E} [f(Z_t)] &\lesssim \sum_{t=2}^T \left(\left(\sum_{q=t-\tilde{m}_t}^{t-1} \Delta_{q+1} \right) + \sqrt{\frac{d}{\tilde{m}_t}} \right) \\ &= \sum_{t=2}^T \min_{m \in \{1, \dots, t-1\}} \left(\left(\sum_{q=t-m}^{t-1} \Delta_{q+1} \right) + \sqrt{\frac{d}{m}} \right). \end{aligned}$$

□

It remains an interesting open problem to determine whether the above guarantee is achievable by a learning rule that has no direct dependence on the Δ_t values: that is, a method that is *adaptive* to variations in the rates of drift. Resolution of this question seems an important step toward applicability of these ideas in practice. Of course, as established in Theorem 1.1, if we instead assume that the asymptotic bound (1.2) holds, then it is possible to replace the direct dependence on Δ_t with a mere dependence on a single parameter α ; however, the price for this is that the finite-sample bound in

Theorem 3.1 would be replaced by an asymptotic guarantee. An alternative option is to suppose the drift rates Δ_t are *bounded* by a value γ , and then provide an algorithm depending only on γ ; this general of a condition on Δ_t precludes the possibility of sublinear cumulative excess risk, but it can nonetheless be interesting to study the dependence of the achieved excess risk on γ . This is the subject of the next subsection.

3.1 Constant Drift Rate

In the context of binary classification, [19, 14, 15, 5, 6, 9, 12] have derived bounds on the sequence of risks (or the number of mistakes) achieved by various methods, under the assumptions that $\{Z_t\}_{t=1}^\infty$ is a product process, and that $\Delta_t \leq \gamma$, for some fixed constant $\gamma \in (0, 1)$. Here we briefly note that some of these results (and in particular, those of [6]) can be generalized to the general learning setting, where we find analogous results on the average of the $\hat{f}_t(Z_t)$ function values. We note that a similar type of result can also be extracted from the analysis of [22].

Let $\bar{m} = \lceil d^{1/3} \gamma^{-2/3} \rceil$. For each integer $t > \bar{m}$, let

$$\bar{f}_t = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{s=t-\bar{m}}^{t-1} f(Z_s).$$

For completeness, for $t \leq \bar{m}$ define \bar{f}_t as an arbitrary element of \mathcal{F} .

Theorem 3.2. *If $\{Z_t\}_{t=1}^\infty$ is a product process, then for $T > 1/\gamma$,*

$$\sum_{t=1}^T \mathbb{E} [\bar{f}_t(Z_t)] - \sum_{t=1}^T \inf_{f \in \mathcal{F}} \mathbb{E} [f(Z_t)] \lesssim (d\gamma)^{1/3} T.$$

We note that a bound similar to that in Theorem 3.2 can in fact be proven for the same method as in Theorem 3.1; indeed, this follows immediately from plugging in γ for the values of Δ_t in the bound, in which case the method itself is also quite similar to the \bar{f}_t in Theorem 3.2. However, as the method described here has a simpler form, we include a brief proof of this result for the stated method.

Proof. As in the proof of Theorem 3.1, the proof is based on the general validity of (2.2). In particular, taking $k_t = 1$ and $m_t = \bar{m} \wedge (t - 1)$, the corresponding \hat{f}_t is equal \bar{f}_t for all $t > \bar{m}$. Thus, (2.2) implies

$$\begin{aligned} \sum_{t=1}^T \mathbb{E} [\bar{f}_t(Z_t)] - \sum_{t=1}^T \inf_{f \in \mathcal{F}} \mathbb{E} [f(Z_t)] &\lesssim \bar{m} + \sum_{t=\bar{m}+1}^T \left(\left(\sum_{q=t-\bar{m}}^{t-1} \Delta_{q+1} \right) + \sqrt{\frac{d}{\bar{m}}} \right) \\ &\leq \bar{m} + \sum_{t=\bar{m}+1}^T \left(\bar{m}\gamma + (d\gamma)^{1/3} \right) \lesssim d^{1/3} \gamma^{-2/3} + (d\gamma)^{1/3} T. \end{aligned}$$

The proof is completed by noting that, for $T > 1/\gamma$, $(d\gamma)^{1/3} T > d^{1/3} \gamma^{-2/3}$, so that $d^{1/3} \gamma^{-2/3} + (d\gamma)^{1/3} T < 2(d\gamma)^{1/3} T$. \square

4 Discussion and Open Problems

As for tightness of the results above, it is not clear whether the rate established in Theorem 1.1 is in any sense optimal. There are several steps in the proof that may introduce slack, but it is not clear whether there is a way to refine the analysis or the method to obtain a smaller exponent in the bound.

A Proof of Lemma 2.2

Since technically the original proof of Lemma 2.2 was stated for identically distributed samples, for completeness we present a brief proof of the result without this restriction. The details follow a standard argument. Specifically, following the usual symmetrization argument (e.g., [7], Lemma 11.4), for (Z''_1, \dots, Z''_m) an independent copy of (Z'_1, \dots, Z'_m) , and $\epsilon_1, \dots, \epsilon_m$ i.i.d. $\text{Uniform}(\{-1, +1\})$ independent of all Z'_i and Z''_i , by Jensen's inequality we have

$$\begin{aligned} \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{t=1}^m (f(Z'_t) - \mathbb{E}[f(Z'_t)]) \right| \right] &= \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{t=1}^m (f(Z'_t) - \mathbb{E}[f(Z''_t)]) \right| \right] \\ &\leq \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{t=1}^m (f(Z'_t) - f(Z''_t)) \right| \right] = \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{t=1}^m \epsilon_t (f(Z'_t) - f(Z''_t)) \right| \right] \\ &\leq 2 \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{t=1}^m \epsilon_t f(Z'_t) \right| \right]. \end{aligned}$$

Then Lemma 6.1 of [21] implies

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \sum_{t=1}^m \epsilon_t f(Z'_t) / \sqrt{m} \right| \middle| Z'_1, \dots, Z'_m \right] \leq 3 \sum_{j=0}^{\infty} 2^{-j} \sqrt{\ln(\mathcal{M}(2^{-j-1}, \mathcal{F}, L_2(P'_m)))},$$

where $\mathcal{M}(\delta, \mathcal{F}, L_p(P'_m))$ is the δ -packing number of \mathcal{F} under $L_p(P'_m)$, and P'_m is the empirical measure induced by Z'_1, \dots, Z'_m . Since functions in \mathcal{F} are bounded in $[0, 1]$, $\mathcal{M}(\delta, \mathcal{F}, L_2(P'_m)) \leq \mathcal{M}(\delta^2, \mathcal{F}, L_1(P'_m))$, and Theorem 6 of [13] (based on Lemma 25 of [23]) implies $\mathcal{M}(\delta^2, \mathcal{F}, L_1(P'_m)) \leq 2 \left(\frac{2e}{\delta^2} \ln \frac{2e}{\delta^2} \right)^d$. Thus, $\sum_{j=0}^{\infty} 2^{-j} \sqrt{\ln(\mathcal{M}(2^{-j-1}, \mathcal{F}, L_2(P'_m)))} \leq c' \sqrt{d}$ for a numerical constant c' . Combining the above inequalities yields the result.

References

- [1] T. M. Adams and A. B. Nobel, *Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling*, Annals of Probability **38** (2010), no. 4, 1345–1367. MR-2663629
- [2] M. Anthony and P. L. Bartlett, *Neural network learning: Theoretical foundations*, Cambridge University Press, 1999. MR-1741038
- [3] P. L. Bartlett, *Learning with a slowly changing distribution*, Proceedings of the 5th Annual Workshop on Computational Learning Theory, 1992, pp. 243–252.
- [4] P. L. Bartlett, S. Ben-David, and S. R. Kulkarni, *Learning changing concepts by exploiting the structure of change*, Machine Learning **41** (2000), 153–174.
- [5] R. D. Barve and P. M. Long, *On the complexity of learning from drifting distributions*, Proceedings of the 9th Conference on Computational Learning Theory, 1996, pp. 122–130. MR-1479321
- [6] ———, *On the complexity of learning from drifting distributions*, Information and Computation **138** (1997), no. 2, 170–193. MR-1479321
- [7] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities*, Oxford University Press, 2013. MR-3185193
- [8] R. C. Bradley, *Absolute regularity and functions of Markov chains*, Stochastic Processes and their Applications **14** (1983), 67–77. MR-0676274
- [9] K. Crammer, Y. Mansour, E. Even-Dar, and J. Wortman Vaughan, *Regret minimization with concept drift*, Proceedings of the 23rd Conference on Learning Theory, 2010, pp. 168–180.
- [10] E. Eberlein, *Weak convergence of partial sums of absolutely regular sequences*, Statistics & Probability Letters **2** (1984), 291–293. MR-0777842
- [11] Y. Freund and Y. Mansour, *Learning under persistent drift*, Proceedings of the 3rd European Conference on Computational Learning Theory, 1997, pp. 109–118. MR-1476925

- [12] S. Hanneke, V. Kanade, and L. Yang, *Learning with a drifting target concept*, Proceedings of the 26th International Conference on Algorithmic Learning Theory, 2015.
- [13] D. Haussler, *Decision theoretic generalizations of the PAC model for neural net and other learning applications*, Information and Computation **100** (1992), 78–150. MR-1175977
- [14] D. P. Helmbold and P. M. Long, *Tracking drifting concepts using random examples*, Proceedings of the 4th Annual Workshop on Computational Learning Theory, 1991, pp. 13–23.
- [15] ———, *Tracking drifting concepts by minimizing disagreements*, Machine Learning **14** (1994), no. 1, 27–45.
- [16] V. Koltchinskii, *Local rademacher complexities and oracle inequalities in risk minimization*, The Annals of Statistics **34** (2006), no. 6, 2593–2656. MR-2329442
- [17] L. Kontorovich, *Measure concentration of strongly mixing processes with applications*, Ph.D. thesis, Carnegie Mellon University, 2007. MR-2710649
- [18] V. Kuznetsov and M. Mohri, *Generalization bounds for time series prediction with non-stationary processes*, Proceedings of The 25th International Conference on Algorithmic Learning Theory, 2014. MR-3295541
- [19] P. M. Long, *The complexity of learning according to two models of a drifting environment*, Machine Learning **37** (1999), no. 3, 337–354. MR-1811576
- [20] Y. Mansour, M. Mohri, and A. Rostamizadeh, *Domain adaptation: Learning bounds and algorithms*, Proceedings of the 22nd Conference on Learning Theory, 2009.
- [21] P. Massart, *Concentration inequalities and model selection. Ecole d’été de Probabilités de Saint-Flour XXXIII - 2003*, Lecture Notes in Mathematics 1896, Springer, 2007. MR-2319879
- [22] M. Mohri and A. Muñoz Medina, *New analysis and algorithm for learning with drifting distributions*, Proceedings of The 23rd International Conference on Algorithmic Learning Theory, 2012.
- [23] D. Pollard, *Convergence of stochastic processes*, Springer-Verlag, Berlin / New York, 1984. MR-0762984
- [24] ———, *Empirical processes: Theory and applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 2, Institute of Mathematical Statistics and American Statistical Association, 1990. MR-1089429
- [25] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, *Learnability, stability and uniform convergence*, Journal of Machine Learning Research **11** (2010), 2635–2670. MR-2738779
- [26] M. Talagrand, *Sharper bounds for gaussian and empirical processes*, The Annals of Probability **22** (1994), 28–76. MR-1258865
- [27] A. W. van der Vaart and J. A. Wellner, *Weak convergence and empirical processes*, Springer, 1996. MR-1385671
- [28] V. Vapnik, *Estimation of dependencies based on empirical data*, Springer-Verlag, New York, 1982. MR-0672244
- [29] ———, *Statistical learning theory*, John Wiley & Sons, Inc., 1998. MR-1641250
- [30] V. Vapnik and A. Chervonenkis, *On the uniform convergence of relative frequencies of events to their probabilities*, Theory of Probability and its Applications **16** (1971), 264–280.
- [31] M. Vidyasagar, *Learning and generalization with applications to neural networks*, 2nd ed., Springer-Verlag, 2003. MR-1938842
- [32] V. A. Volkonskii and Y. A. Rozanov, *Some limit theorems for random functions. I*, Theory of Probability and its Applications **4** (1959), 178–197. MR-0121856
- [33] L. Yang, *Active learning with a drifting distribution*, NIPS, 2011.
- [34] B. Yu, *Rates of convergence for empirical processes of stationary mixing sequences*, The Annals of Probability **22** (1994), no. 1, 94–116. MR-1258867